Shape Driven Kernel Adaptation in CNN for Robust Facial Trait Recognition

Shaoxin Li[1,3], Junliang Xing[2,3], Zhiheng Niu[3], Shiguang Shan[1], Shuicheng Yan[3]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[2]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China.
[3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore

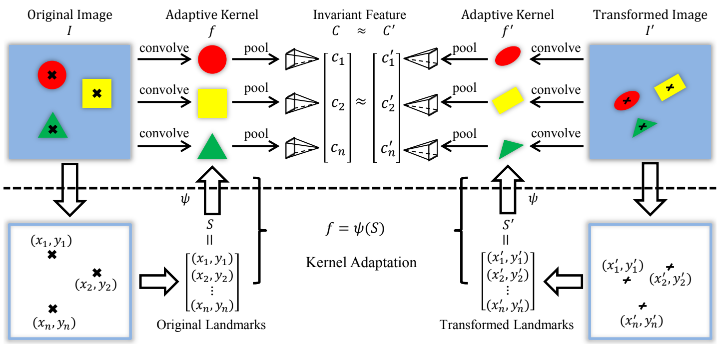

Figure 1: An toy example of kernel adaptation in the CNN framework.

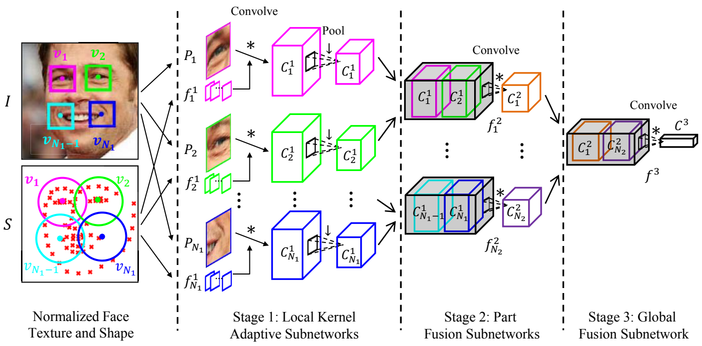

Figure 2: Flowchart of the tree-structured kernel adaptive CNN.

One key challenge of facial trait recognition is the large non-rigid appearance variations due to some irrelevant real world factors, such as viewpoint and expression changes. Current Convolutional Neural Network (CNN) based methods learn discriminant features mainly from texture information thus suffers from these real world nuisance variables. Although state-of-the-art deep CNN models [1, 4] are proven to be powerful in handling these complex factors and learning invariant features, however very deep and large network structure seems to be essential to achieve such invariance [4].

Rather than using deeper and larger networks, in this paper, we explore how the shape information, i.e. facial landmark positions, can be explicitly deployed into the popular CNN architecture to learn invariant features in a more intuitive and compact way.

First, instead of using fixed kernels, we propose a kernel adaptation method to dynamically determine the convolutional kernels according to the positions of facial landmarks S, as shown in expression (1).

$$f = \psi(S, \Theta), \tag{1}$$

where $\psi(\cdot)$ is an adaptation function that can depict the relationship between the facial landmarks S and the proper kernel f. Θ is the parameter of ψ. A sketch of the basic idea is shown in Figure 1. As aforementioned, due to real world variation, the appearance of an image I may be significantly different to its transformed version of image I'. However, if proper kernel adaptation function $\psi(\cdot)$ is learned to generate a transformed version of kernel (also called convolutional filter), the convolutional feature maps would become invariant to these transformations.

Although the ideal adaptation function ψ may be very complex, in this paper, we use a simple linear function to approximate it. Formally, this liner function in our kernel adaptation method can be represented as:

$$f = W \cdot S, \tag{2}$$

where W is the linear matrix used to generate the adaptive kernel f. With kernel adaptation as indicated by Eqn. (2), given an input face image I, the kernel functions f can be adaptively generated according to its shape information S. As a result, the feature learning process can automatically achieve certain complex geometric transformation invariance.

Second, motivated by the intuition that appearance variation caused by pose and expression is non-rigid, different facial components may demand different kernel adaptation functions. Therefore, instead of using single adaptation function over the whole face, the kernel adaption is separately adopted in multiple local CNN subnetworks, indicated as C_i $(i = 1, 2, ..., N)$, over multiple local facial patches, indicated as P_i $(i = 1, 2, ..., N)$. In this way, each small facial patch P_i has its own adaptation function W_i. Moreover, only landmarks around the patch P_i contain valuable information for modeling the appearance deformations in this local patch. Thus, we only use local shape information S_i to infer the local adaptive kernel f_i of the local patch P_i. Formally, for each local subnetwork C_i, we represent its adaptive kernel f_i as a function of corresponding "shape" information S_i:

$$f_i = W_i \cdot S_i, \tag{3}$$

As the variation caused by pose and expression in each small local patch can be assumed as a rigid transformation approximately. This local linear adaptation function is capable in depicting the relation between local shape information and desired kernel function.

To jointly learn features from multiple local regions, we further propose a tree-structured convolutional architecture to hierarchically fuse multiple local adaptive CNN subnetworks. As shown in Figure 2, given a normalized face image I and corresponding facial landmarks $S = \{v_i\}_{i=1}^{N_1}$, multiple local kernel adaptive CNN subnetworks $\{C_i^1\}_{i=1}^{N_1}$ are constructed to learn features from multiple local patches $\{P_i\}_{i=1}^{N_1}$. The convolved features learned by multiple local subnetworks are then combined as the middle-level representations to learn high-level features with the fusion subnetworks, i.e. multiple part fusion subnetworks $\{C_i^2\}_{i=1}^{N_2}$ and a global fusion subnetwork C^3. Finally, a logistic regression layer is used to generate the final prediction y from vectorized global convolved features X. The whole networks can be trained with common back-propagation method [2] in the end-to-end manner.

Implementations and comparisons are detailed in the paper. Demonstrated in the experiments, own to the usage of kernel adaptation, with relatively shallow networks and much less parameters, our method can achieve comparable or better performance compared to state-of-the-art deep learning methods [1, 3] and other facial trait recognition methods.

[1] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[2] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 1998.

[3] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[4] Matthew Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*. 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.